NIST GenAI

https://ai-challenges.nist.gov/genai

July 8, 2024

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

INFORMATION TECHNOLOGY LABORATORY

NIST GenAI Team

- Yooyoung Lee (PI)
- Seungmin Seo (Computer Scientist, FGR)
- Lukas Diduch (Research Engineer, CTR)
- George Awad (Computer Scientist)
- Kay Peterson (Social Scientist)
- Hari Iyer (Statistician/Forensic Statistician)

Acknowledgements

- Advisory
  - Mark Przybocki, Division Chief, IAD
  - Ian Soboroff, Retrieval Group, IAD
  - Jim Horan, Multimodal Information Group, IAD
- Peter Fontana: GenAI Code (coming soon)
- Baptiste Chocot: Human Survey website development

Outline

- Webinar Logistics
- NIST GenAI Program
- NIST ARIA Program
- GenAI Pilot: Text-to-Text (T2T)
- GenAI Pilot: Text-to-Image (T2I)
- GenAI Future Directions
- GenAI Discussion Session

Logistics

- We encourage the use of Chat and Q&A throughout the meeting
- Please mute during the Presentation portion
  - Moderators may mute participants if needed (eg. eliminate background noise). If muted by mistake, please unmute yourself
- There will be a “Discussion” Section
  - Please raise hand if you desire to unmute.
  - Continue to use Chat and Q&A
- Meeting will not be recorded; Chat, Q&A, and Internal notes will be saved.

Two New AI Programs under NIST ITL

- **NIST GenAI (Evaluating Capabilities & Limitations of Generative AI & Discriminative AI Technologies)**

  https://ai-challenges.nist.gov/genai

  Yooyoung Lee (PI)

- **NIST ARIA (Assessing Risks and Impacts of AI)**

  https://ai-challenges.nist.gov/aria

  Reva Schwartz (PI)

Can Your Model Handle This?
Testing in the Real World

NIST's Assessing Risks and Impacts of AI (ARIA)

NIST
NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

ARIA expands the scope of evaluations to include **people and how they use AI** in real world conditions.

ARIA's three evaluation levels, can provide more direct knowledge about:

how AI capabilities (in model testing) can connect to risks (in red teaming) and positive and negative impacts (in field testing)

ARIA's three testing levels

Model Testing

Claimed model capabilities.

Red Teaming

Adverse outcomes and **how** they occur.

Model guardrails.

Field Testing

Positive and negative impacts of AI under **regular use.**

Short term insights:

- functionality across risks and contexts
- effectiveness of guardrails and mitigations
- test applicability for each risk

Long-term outcomes:

- guidelines
- tools
- evaluation methods
- metrics

ARIA will advance our understanding of AI’s negative and positive impacts to people and society.

ARIA is like other NIST community evaluations.

- Designed to improve evaluation state of practice focused on a key challenge.
- Builds up a dedicated research community.
- Open to all, teams opt-in to participate.
- Evaluation output is made available for future research.

ARIA 0.1 will be a pilot evaluation focused on **risks and impacts** of large language models (LLMs).

ARIA is **not** designed to test AI systems for operational, reporting, certification or oversight purposes.

NIST GenAI

https://ai-challenges.nist.gov/genai

Human or AI?

Generative AI

NIST Unveils Innovative Tool to Assess Generative AI

Analytics Insight

Credit source: https://www.analyticsinsight.net/generative-ai/nist-unveils-innovative-tool-to-assess-generative-ai

*“**NIST's innovative tool**'s first venture is a pilot to construct frameworks that can dependably say the contrast between human-created and AI-generated media, beginning with content. Whereas numerous administrations imply identifying deepfakes, studies and their testing have appeared to be unstable at best, especially when it comes to content. NIST GenAI is inviting groups from the scholarly world and industry to inquire about labs to yield either generators, AI systems to produce substance, or discriminators, which are frameworks planned to distinguish AI-generated content…”*

*“NIST <u>Generative AI</u>'s launch and deep fake-focused study come as the volume of AI-generated deception and disinformation information grows exponentially. **NIST's innovative tool** promises to revolutionize Generative AI models…”*

"He just took random photos from my daughter's from prom and turned them into nude images and started distributing them among student body," XXX's mother said.
https://wgntv.com/far-north-suburbs/investigation-into-30-explicit-ai-generated-photos-of-suburban-high-school-students-underway/

The rise of AI fake news is creating a 'misinformation superspreader' *AI is making it easy for anyone to create propaganda outlets, producing content that can be hard to differentiate from real news.*
https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

The People Onscreen Are Fake. The Disinformation Is Real. *In one video, a news anchor with perfectly combed dark hair and a stubbly beard outlined what he saw as the United States' shameful lack of action against gun violence.*
https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

"Medical AI could be 'dangerous' for poorer nations, WHO warns" *The rapid growth of generative AI in healthcare has prompted the agency to set out guidelines for ethical use.*
https://www.nature.com/articles/d41586-024-00161-1

"She was accused of faking an incriminating video of teenage cheerleaders. She was arrested, outcast and condemned. The problem? Nothing was fake after all"
https://www.theguardian.com/technology/article/2024/may/11/she-was-accused-of-faking-an-incriminating-video-of-teenage-cheerleaders-she-was-arrested-outcast-and-condemned-the-problem-nothing-was-fake-after-all

What is NIST GenAI?

- An umbrella program that supports various evaluations for research in Generative AI in different modalities (text, image, video, audio, code)
  - Measure and understand the capabilities and limitations of cutting-edge generative AI technologies
  - Develop performance metrics and conduct comparative analysis of different AI systems using relevant metrics,
  - Create evolving benchmark datasets in generative adversarial framework,
  - Bridge a gap between research and real-world scenarios,
  - Help stakeholders (e.g., government, private sector, and academia) develop approaches for ensuring the trustworthiness of information, and
  - Promote information integrity and guide responsible use of digital content.

What’s Unique about NIST GenAI?

- Support various evaluation series in different modalities
- A generative adversarial (cat-and-mouse) test framework
- A parallel comparison of human and AI evaluation paradigms
  - Human assessments via surveys
    - Human generators
    - Human discriminators
  - AI system evaluations via relevant metrics
    - AI generators
    - AI discriminators
  - Compare human performance with AI system performance

*“NIST Generative AI's launch…**NIST's innovative tool** promises to revolutionize Generative AI …”,news article*

Text

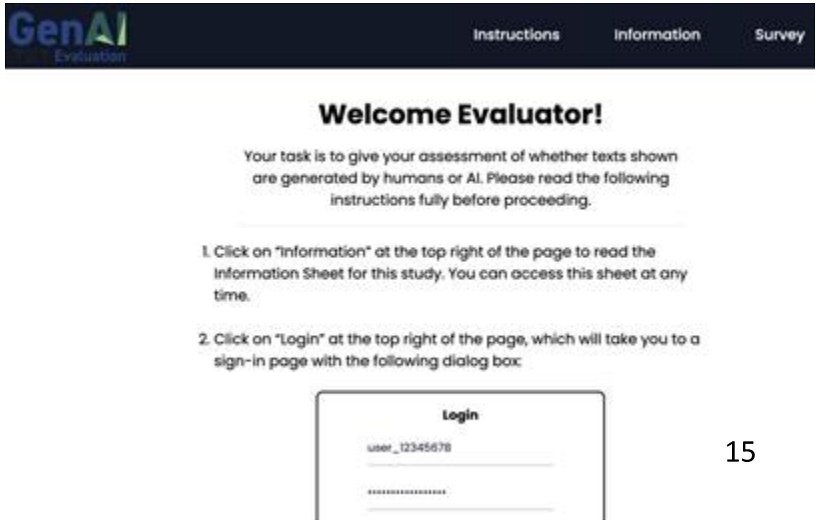
Image

Video

Voice

Code

Multimodal



Welcome Evaluator!

Your task is to give your assessment of whether texts shown are generated by humans or AI. Please read the following instructions fully before proceeding.

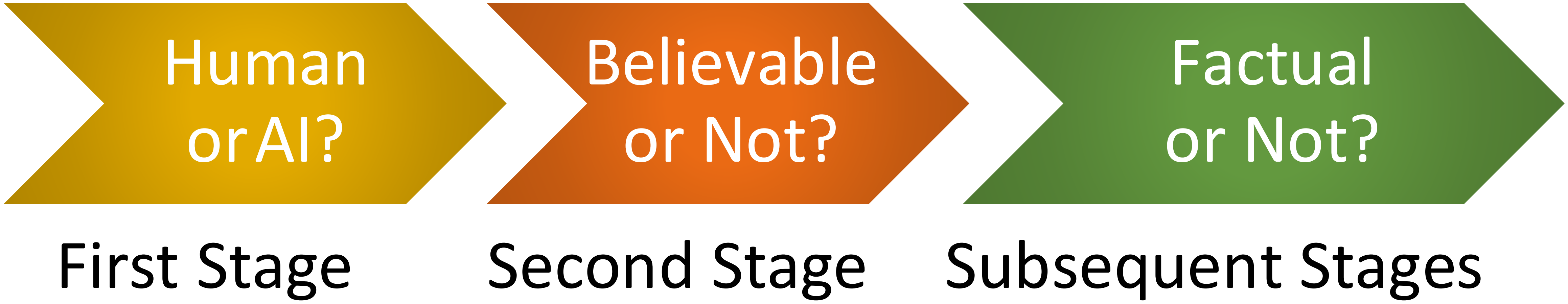
1. Click on "Information" at the top right of the page to read the Information Sheet for this study. You can access this sheet at any time.
2. Click on "Login" at the top right of the page, which will take you to a sign-in page with the following dialog box:

Login

user_12345678

••••••••••••••••

NIST GenAI in Multiple Stages



Human or AI?

Believable or Not?

Factual or Not?

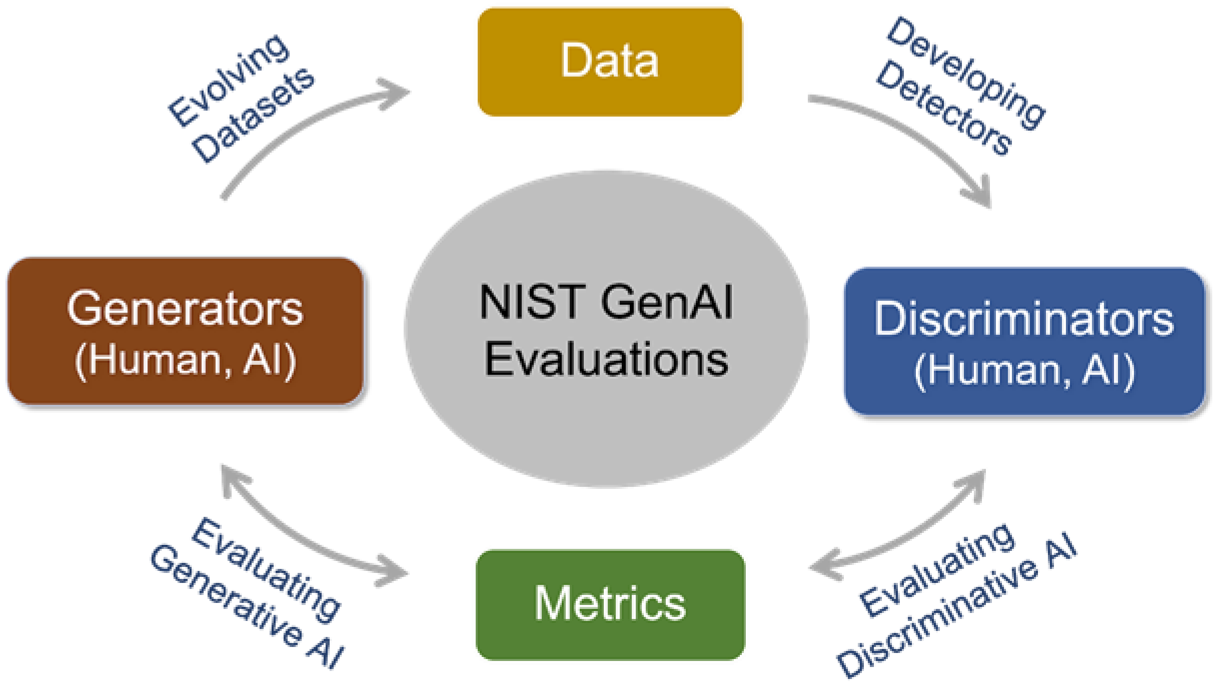
First Stage

Second Stage

Subsequent Stages

- First stage (pilot): "indistinguishability” -> Human, AI, (Human+AI)
- Second stage: "indistinguishability” + "plausibility (believability)”
- Subsequent stages: will decide based on lessons learned from the previous stages and feedback from participants and stakeholders (e.g., factual, confabulation, or disinformation)

NIST GenAI “Pilot”



- The initial effort aims to develop a robust evaluation pipeline that can be used to learn the capabilities of *generative AI & discriminative AI* systems focusing on “text”
- Two types of participants using a generative adversarial test framework
  - Generators: generate evolving datasets
    - **AI generators: automatically generate synthetic content that is indistinguishable from human-produced content**
    - Human generators: manually create authentic content
  - Discriminators: detect synthetic content
    - **AI Discriminators: automatically detect synthetic content**
    - Human Discriminators: manually detect synthetic content

GenAI Pilot: Text-to-Text (T2T)

- T2T Generators (T2T-G) task
  - automatically generate high-quality summaries given a statement of information needed ("topic") and a set of source documents to summarize
- T2T Discriminators (T2T-D) task
  - automatically detect if a target summary has been generated using a Generative AI system or a Human

Summary Example:
AI charged the government of Kenya in 1997 with police torture of suspects and footdragging on promised human rights reforms. The Kenyan government complained that AI must be sensitive to national sovereignty and the human rights of killed police officers and accused AI of inciting Kenyans against the government. A 1998 AI report charged the Rwandan army with massacring hundreds of unarmed civilians. Rwandan authorities countered that AI was "another hand of the insurgency" and a mouthpiece for the Hutu hardliners. AI charges of racist acts by the German police were thoroughly investigated by German state officials in what they termed an impartial review that found that the police were not racist as a whole. AI complaints against the Afghan Taliban in 1996 were deemed interference in the internal affairs of an Islamic state.

GenAI Metrics: Generator

- Measures of the extent to which the generated summary makes use of source articles provided by NIST

- Each summary will be compared separately to 25 source articles on a given topic.

- As Round 1 contains 10 topics, each metric score displayed on the leaderboard will represent the average of 250 scores (10 summaries x 25 source articles per summary)

- Each summary should pass toxicity checks. Participants will be notified via automatic email if the submitted summary contains toxic content (e.g., hateful, aggressive, rude, unreasonable or disrespectful comment).

GenAI Metrics: Generator

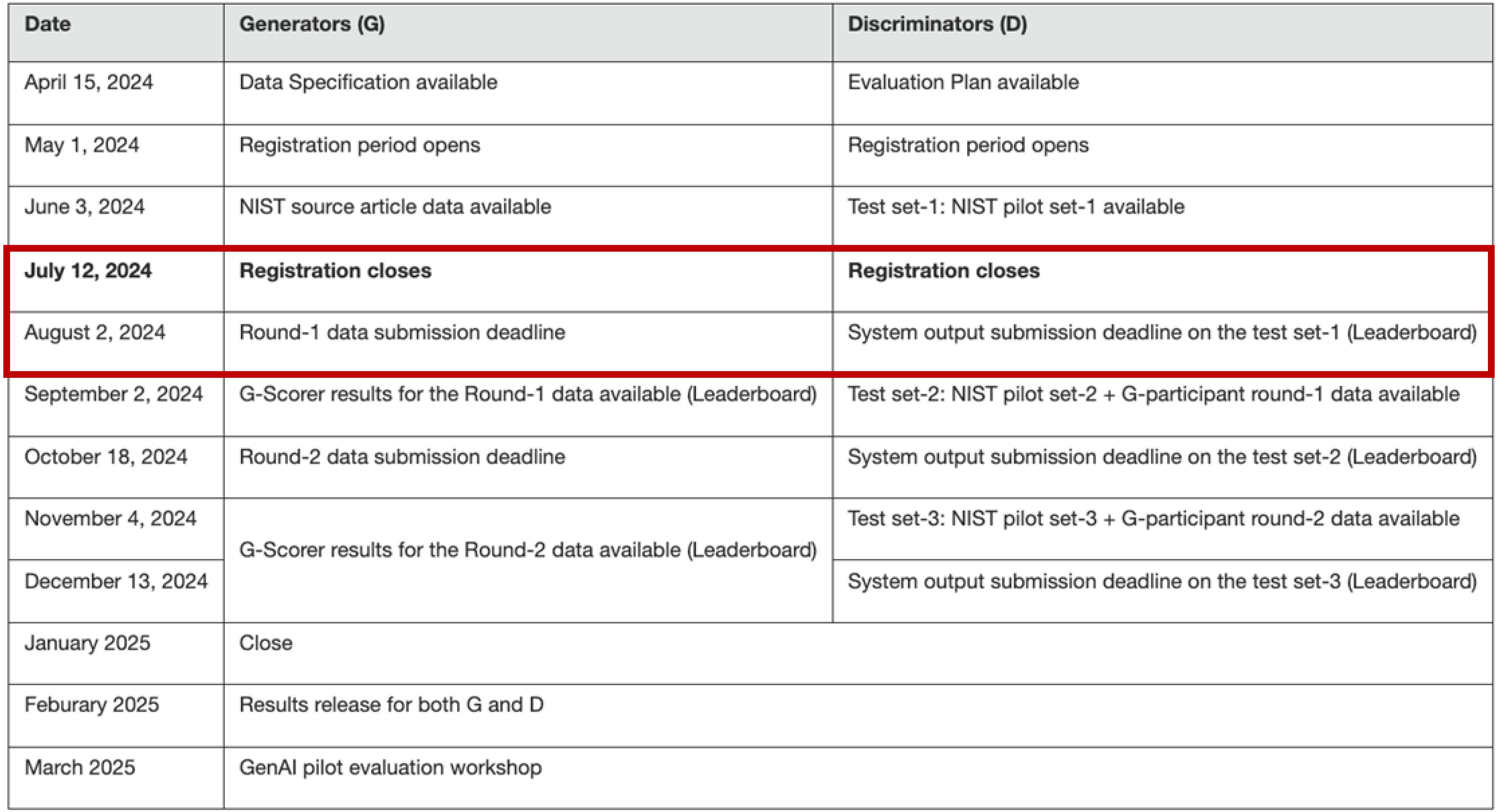
- Discriminator_Max_AUC: A higher value is considered better. It will not be available until running discriminators on a generator team's data.

- LLM_detector_n scores: Confidence score of NIST's baseline discriminator. Higher scores from each detector indicate that NIST's baseline discriminator judges the given summary to be AI-generated content. 0 (human-written) ~ 1 (AI-generated)

GenAI Metrics: Discriminator

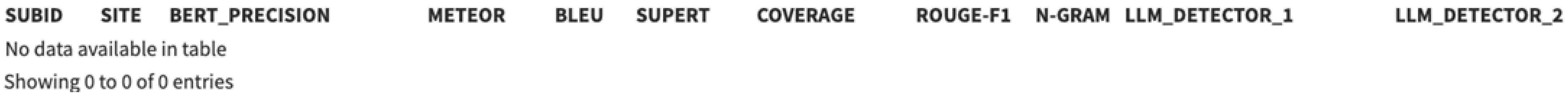
- AUC (Area Under roc Curve)
- EER (Equal Error Rate)
- AUC@FPR=0.1
- TPR@FPR=0.1
- TNR@FNR=0.1
- ROC and DET plots (provided to participants)
- Brier Score (Discrimination + Calibration) – coming soon
- Please include, in your system output filename, the recommended ***cutoff (threshold)*** for the confidence scores for binary classification.

GenAI T2T Schedule



Date	Generators (G)	Discriminators (D)
April 15, 2024	Data Specification available	Evaluation Plan available
May 1, 2024	Registration period opens	Registration period opens
June 3, 2024	NIST source article data available	Test set-1: NIST pilot set-1 available
July 12, 2024	**Registration closes**	**Registration closes**
August 2, 2024	Round-1 data submission deadline	System output submission deadline on the test set-1 (Leaderboard)
September 2, 2024	G-Scorer results for the Round-1 data available (Leaderboard)	Test set-2: NIST pilot set-2 + G-participant round-1 data available
October 18, 2024	Round-2 data submission deadline	System output submission deadline on the test set-2 (Leaderboard)
November 4, 2024	G-Scorer results for the Round-2 data available (Leaderboard)	Test set-3: NIST pilot set-3 + G-participant round-2 data available
December 13, 2024		System output submission deadline on the test set-3 (Leaderboard)
January 2025	Close	
Feburary 2025	Results release for both G and D	
March 2025	GenAI pilot evaluation workshop	

GenAI T2T Generator Leaderboard

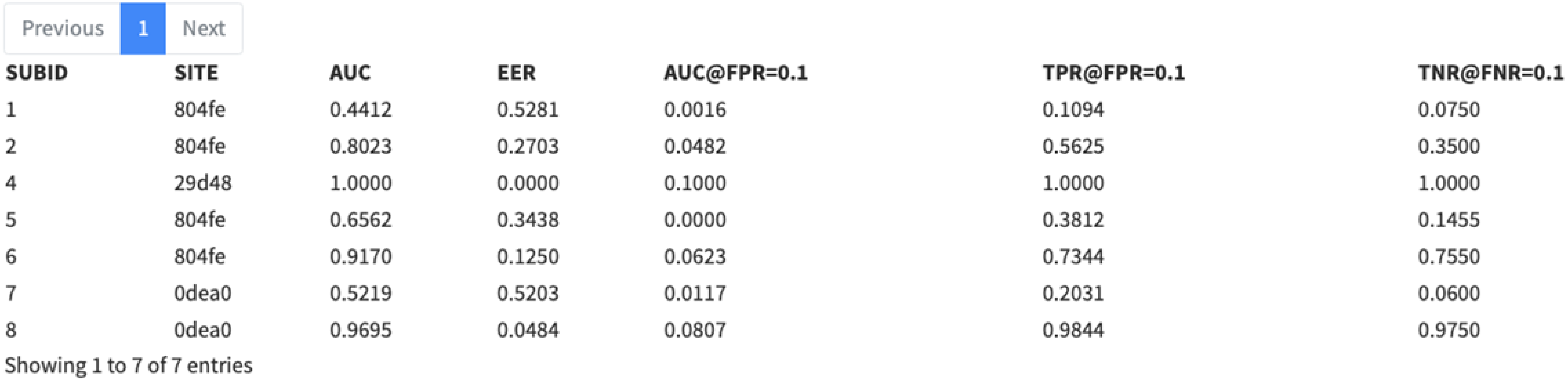


SUBID	SITE	BERT_PRECISION	METEOR	BLEU	SUPERT	COVERAGE	ROUGE-F1	N-GRAM	LLM_DETECTOR_1	LLM_DETECTOR_2

No data available in table

Showing 0 to 0 of 0 entries

T2T Discriminator Leaderboard



SUBID	SITE	AUC	EER	AUC@FPR=0.1	TPR@FPR=0.1	TNR@FNR=0.1
1	804fe	0.4412	0.5281	0.0016	0.1094	0.0750
2	804fe	0.8023	0.2703	0.0482	0.5625	0.3500
4	29d48	1.0000	0.0000	0.1000	1.0000	1.0000
5	804fe	0.6562	0.3438	0.0000	0.3812	0.1455
6	804fe	0.9170	0.1250	0.0623	0.7344	0.7550
7	0dea0	0.5219	0.5203	0.0117	0.2031	0.0600
8	0dea0	0.9695	0.0484	0.0807	0.9844	0.9750

Showing 1 to 7 of 7 entries

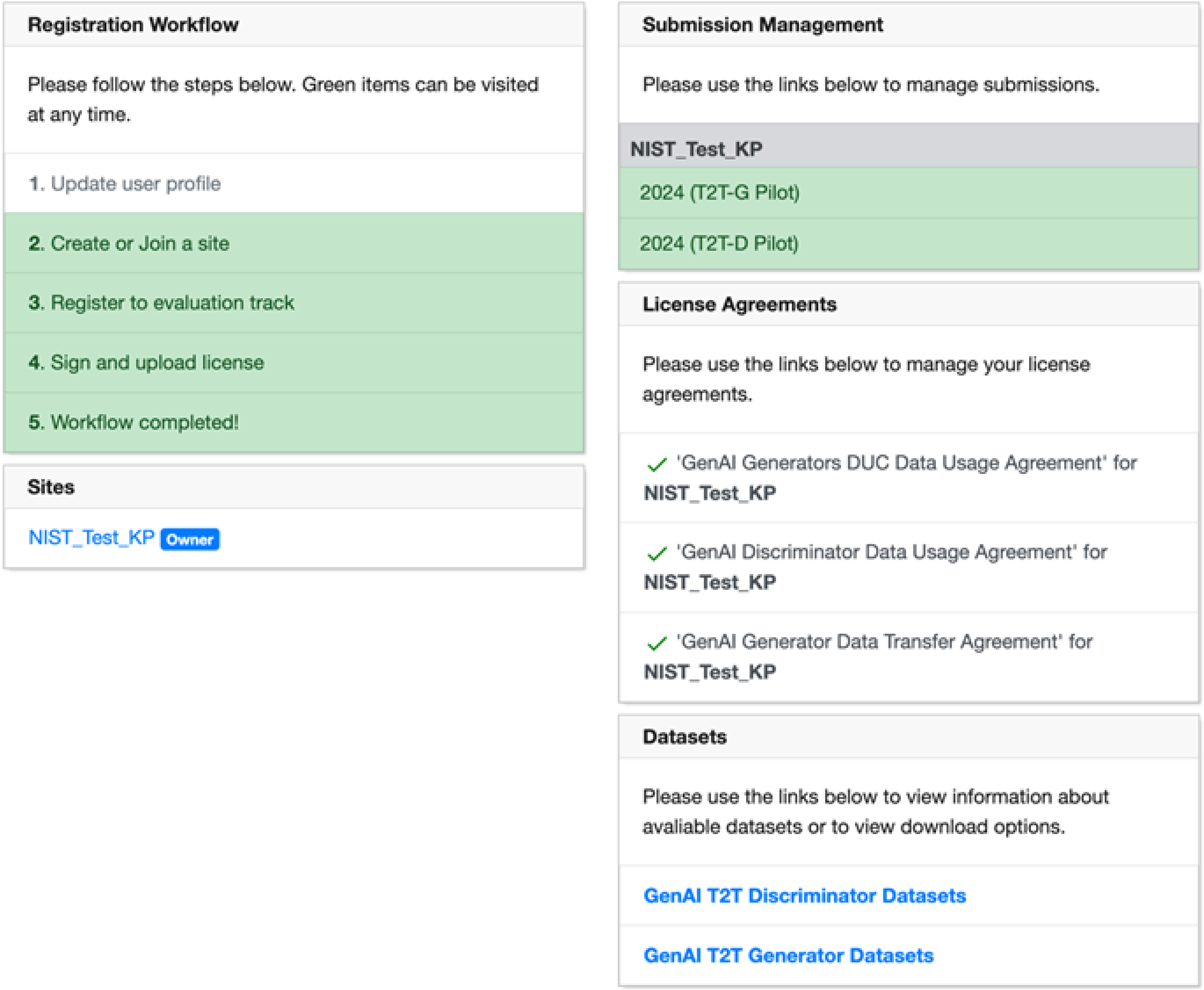
We emphasize that this is a pilot study (Round-1 Submissions on Testset-1). **The primary purpose of the GenAI pilot is to develop an evaluation pipeline between the NIST team and participants.** Therefore, we encourage all participants to submit their system output, regardless of their system performance.

GenAI Pilot: Text-to-Image (T2I)
(Coming Soon)

- T2I Generators (T2I-G) task
  - automatically generate high-quality mugshots and artwork-like images
- T2I Discriminators (T2I-D) task
  - automatically detect if a target image has been generated using a Generative AI system or a Human (e.g., a photo of a real person or real artwork)

NIST GenAI Registration

- Start with login.gov account.
- Log in at https://ai-challenges.nist.gov/genai.
- From your dashboard, follow the registration workflow steps 1 -5.
- Deadline for submitting all required registration documentation to NIST: **July 12, 2024**



Registration Workflow

Please follow the steps below. Green items can be visited at any time.

1. Update user profile
2. Create or Join a site
3. Register to evaluation track
4. Sign and upload license
5. Workflow completed!

Sites

NIST_Test_KP Owner

Submission Management

Please use the links below to manage submissions.

NIST_Test_KP

2024 (T2T-G Pilot)

2024 (T2T-D Pilot)

License Agreements

Please use the links below to manage your license agreements.

✓ 'GenAI Generators DUC Data Usage Agreement' for NIST_Test_KP

✓ 'GenAI Discriminator Data Usage Agreement' for NIST_Test_KP

✓ 'GenAI Generator Data Transfer Agreement' for NIST_Test_KP

Datasets

Please use the links below to view information about avaliable datasets or to view download options.

GenAI T2T Discriminator Datasets

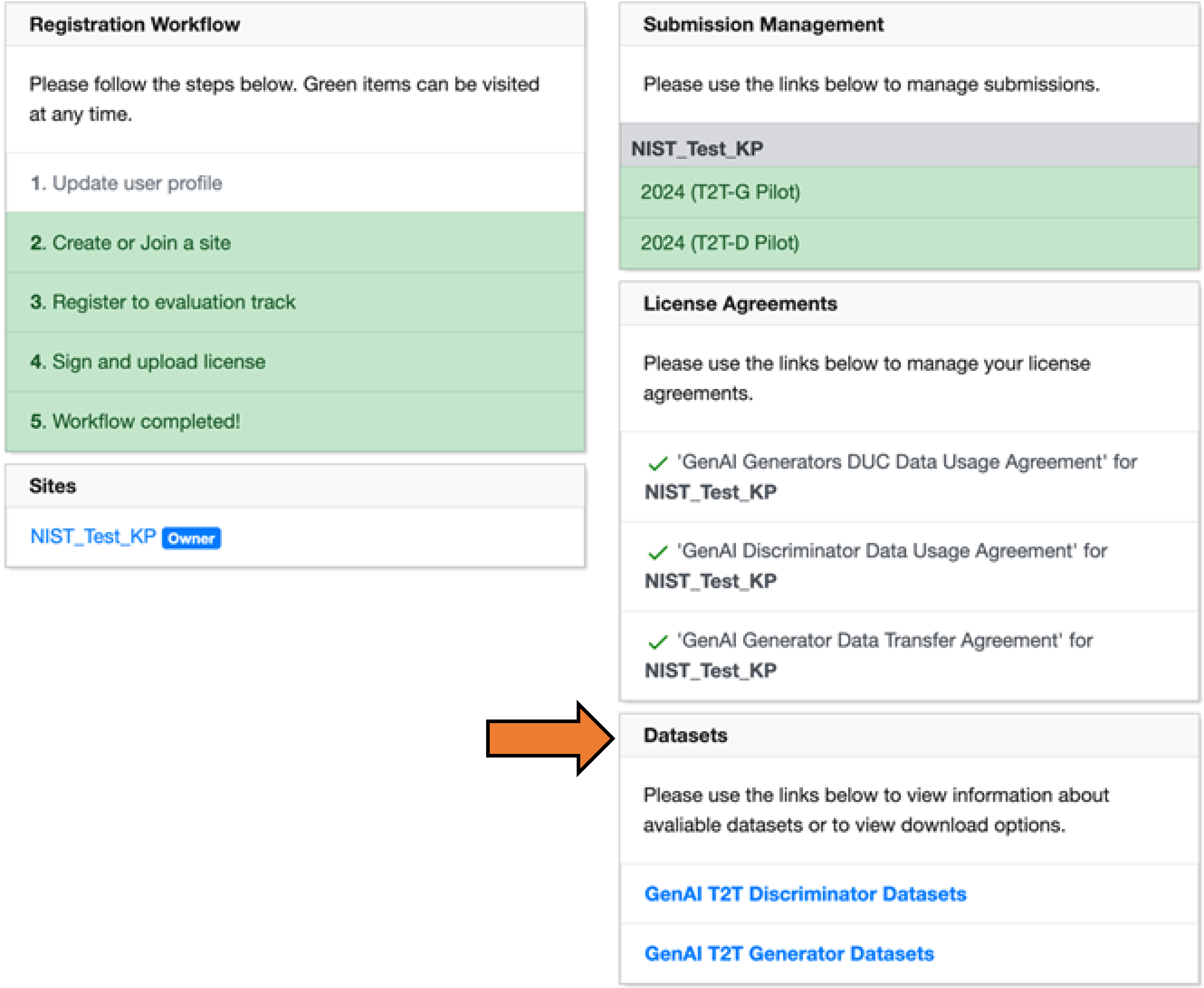
GenAI T2T Generator Datasets

NIST GenAI Registration - Workflow

1. Complete all fields of user profile.
    - ***Individuals can only participate on behalf of an organization; fill the Affiliation field accordingly.***
2. Create or join site.
3. One user per site: Register the site for task(s): Generator, Discriminator, or both.
    - ***Do not skip this step.***
4. One user per site: Complete and submit license agreements based on selections in step 3:
    - **Generator - *2 agreements required*:**
        - Generator data usage agreement (download from website, complete and upload via website)
        - Generator data transfer agreement (download from website, complete and return ***via email*** to agreements@nist.gov, cc'ed to genai-poc@nist.gov). ***Do not modify the wording of the agreement. Please make sure this agreement is completed by someone authorized to sign DTAs on behalf of your organization.***
    - **Discriminator - *1 agreement required*:**
        - Discriminator data usage agreement (download from website, complete and upload via website)
    - Use License agreements section of dashboard to keep track of agreements.
5. Done!

NIST GenAI Registration - Approval and Data

- Approval timeline after completed registration:
  - Discriminator: 1-2 days
  - Generator: 1-2 weeks
- Approval email will go to registered email address
- Evaluation data becomes available in the Datasets section



Registration Workflow

Please follow the steps below. Green items can be visited at any time.

1. Update user profile
2. Create or Join a site
3. Register to evaluation track
4. Sign and upload license
5. Workflow completed!

Sites

NIST_Test_KP Owner

Submission Management

Please use the links below to manage submissions.

NIST_Test_KP

2024 (T2T-G Pilot)

2024 (T2T-D Pilot)

License Agreements

Please use the links below to manage your license agreements.

✓ 'GenAI Generators DUC Data Usage Agreement' for NIST_Test_KP

✓ 'GenAI Discriminator Data Usage Agreement' for NIST_Test_KP

✓ 'GenAI Generator Data Transfer Agreement' for NIST_Test_KP

Datasets

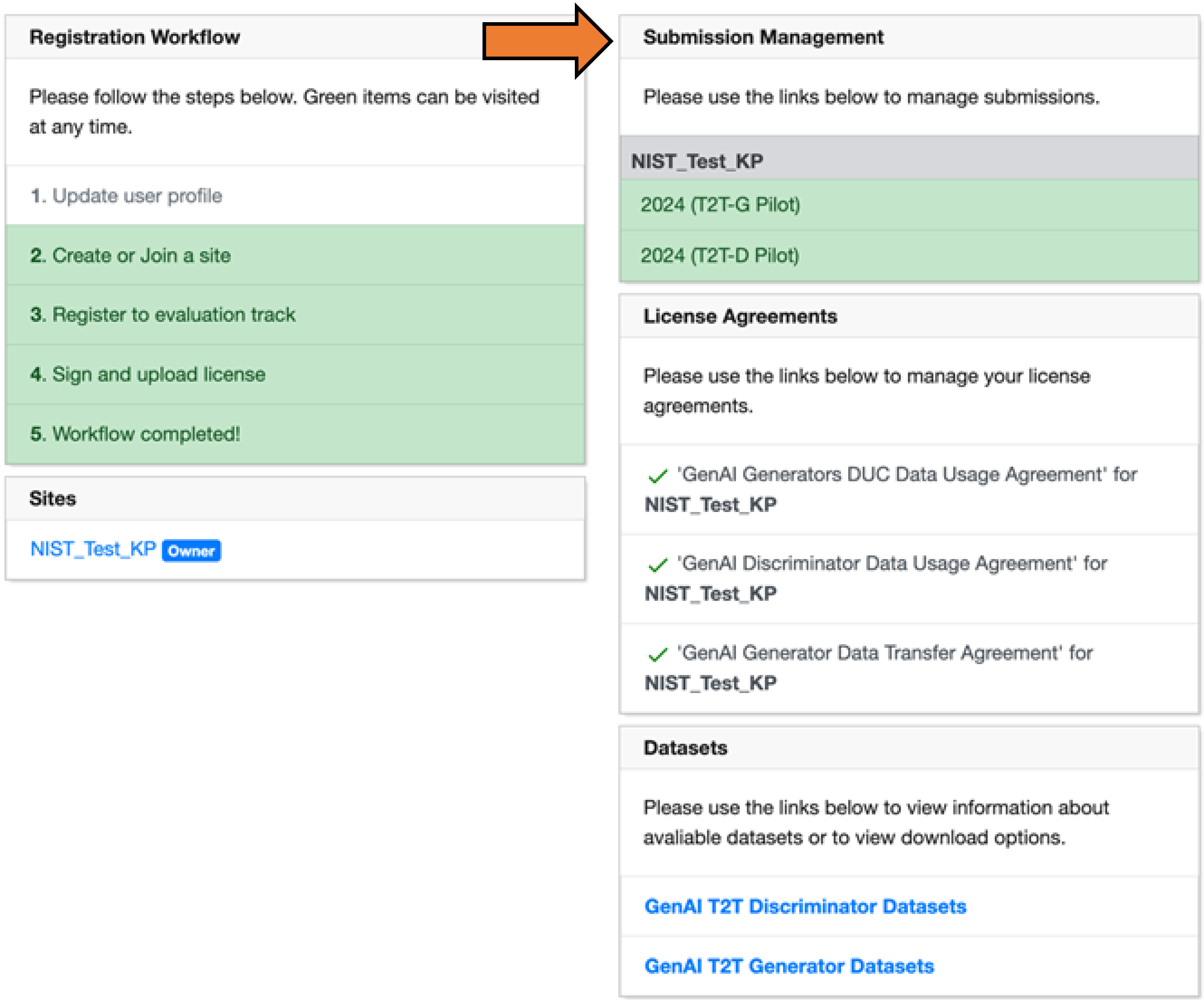
Please use the links below to view information about avaliable datasets or to view download options.

GenAI T2T Discriminator Datasets

GenAI T2T Generator Datasets

NIST GenAI - Making Submissions

- Use the Submission Management section of your dashboard to make submissions.
- The evaluation plan for each task has detailed instructions for submission formatting and validation.



Registration Workflow

Please follow the steps below. Green items can be visited at any time.

1. Update user profile

2. Create or Join a site

3. Register to evaluation track

4. Sign and upload license

5. Workflow completed!

Sites

NIST_Test_KP Owner

Submission Management

Please use the links below to manage submissions.

NIST_Test_KP

2024 (T2T-G Pilot)

2024 (T2T-D Pilot)

License Agreements

Please use the links below to manage your license agreements.

✓ 'GenAI Generators DUC Data Usage Agreement' for NIST_Test_KP

✓ 'GenAI Discriminator Data Usage Agreement' for NIST_Test_KP

✓ 'GenAI Generator Data Transfer Agreement' for NIST_Test_KP

Datasets

Please use the links below to view information about avaliable datasets or to view download options.

GenAI T2T Discriminator Datasets

GenAI T2T Generator Datasets

NIST GenAI Future Directions

- NIST GenAI will consider “Believability”
- NIST GenAI Code challenge (Coming Soon)
  - Q: Can AI-generated code be used effectively in testing software?
- NIST GenAI Voice challenge
  - e.g., Assess synthetic (human-like) speech audio and its detection
- NIST GenAI Deepfake (Video&Voice) challenge
- NIST GenAI Forensics

Q & A

We welcome any questions, feedback, ideas, or suggestions:

- Program goals, objectives and long term vision
- Team registration and participation
- G- and D-submissions (How can the GenAI team assist you better for easier participation?)
- Evaluation workflow
- Datasets, tasks, modalities, etc
- Workshop participation (G- and D-submitted participants only)
- Collaboration ideas (across teams, teams + NIST)

Contact: genai-poc@nist.gov

Thank you!

https://ai-challenges.nist.gov/genai

GenAI: genai-poc@nist.gov

https://ai-challenges.nist.gov/aria

ARIA: aria_inquiries@nist.gov

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

ITL INFORMATION TECHNOLOGY LABORATORY